NDT&E International 154 (2025) 103392

Contents lists available at ScienceDirect

NDT and E International

journal homepage: www.elsevier.com/locate/ndteint

Human-machine collaborative automation strategies for ultrasonic phased array data analysis of carbon fibre reinforced plastics

Vedran Tunukovic [a,b,*], Shaun McKnight [a], Amine Hifi [a], Ehsan Mohseni [a], S. Gareth Pierce [a], Randika K.W. Vithanage [a], Gordon Dobie [a], Charles N. MacLeod [a], Sandy Cochran [b], Tom O'Hare [c]

[a] Sensor Enabled Automation, Robotics, and Control Hub (SEARCH), Centre for Ultrasonic Engineering (CUE), Electronic and Electrical Engineering Department, University of Strathclyde, Glasgow, UK
[b] Future Ultrasonic Engineering, FUSE CDT, Glasgow, UK
[c] Spirit AeroSystems, Belfast, UK

ARTICLE INFO

Keywords:
Automated phased array inspection and data interpretation
Machine learning for phased array ultrasonic testing
Non-destructive evaluation of aerospace composites
Carbon fibre reinforced polymers
Multi model machine learning for full non-destructive evaluation automation

ABSTRACT

NDE 4.0 represents the integration of recent advancements in robotics, sensor technology, and Artificial Intelligence (AI), transforming and automating traditional NDE in line with Industry 4.0 principles. Despite these advancements, data analysis in NDE is still largely performed manually or with traditional rule-based tools such as signal thresholding. These tools often struggle to effectively manage complex data patterns or high noise levels, leading to unreliable defect detection. Additionally, they require frequent manual adjustments to set appropriate parameters for varying inspection conditions, which can be inefficient and error-prone in dynamic or fast paced environments. In contrast, AI-based analysis tools have demonstrated improvements over traditional methods, offering greater accuracy in defect detection and adaptability to higher variability within captured signals. However, their adoption in industrial settings remains limited due to challenges associated with model trust and their "black box" nature. Additionally, practical guidelines for implementing AI tools into NDE workflow are rarely discussed, motivating this work to explore various integration strategies across different automation levels. Three levels of automation were explored, ranging from basic AI-assisted workflows, where tools provide suggestions, to advanced applications where multiple AI models simultaneously process data in a comprehensive analysis, shifting human operators to a supervisory role. Proposed strategies of AI integration into the NDE automation workflow were evaluated on inspection of two defective complex-geometry carbon fibre-reinforced plastics components, commonly used in aerospace and energy sectors for safety-critical structures such as aircraft fuselages and wind turbine blades. The experimental scans were conducted using a phased array ultrasonic testing roller probe mounted on an industrial manipulator, closely replicating industrial practices, and successfully identifying 36 manufactured defects through a combination of supervised object detection on amplitude C-scans, unsupervised anomaly detection on ultrasonic B-scans, and a self-supervised AI model for processing full volumetric ultrasonic data. This inclusion of multiple AI models led to an improvement of up to 17.2 % in the F1 score compared to single-model approaches. Unlike manual inspections, which take hours for larger components, the proposed approach completes the analysis in 94.03 and 57.01 s for the two inspected samples, respectively.

1. Introduction

Non-Destructive Evaluation (NDE) encompasses techniques for inspecting materials without altering their properties. Common methods include radiographic testing, thermographic testing, electromagnetic inspection, and Ultrasonic Testing (UT). NDE can be performed at multiple stages throughout the lifecycle of a material or component: during manufacturing (both in-process and post-manufacturing), throughout its service life with periodic inspections, and after decommissioning at end-of-life. NDE can be broadly divided into three stages:

* Corresponding author;Sensor Enabled Automation, Robotics, and Control Hub (SEARCH), Centre for Ultrasonic Engineering (CUE), Electronic and Electrical Engineering Department, University of Strathclyde, Glasgow, UK
E-mail address: vedran.tunukovic@strath.ac.uk (V. Tunukovic).

https://doi.org/10.1016/j.ndteint.2025.103392
Received 17 December 2024; Received in revised form 14 February 2025; Accepted 15 March 2025
Available online 16 March 2025

sensor delivery, data collection, and data interpretation. Traditionally, sensor delivery and data collection in NDE processes were performed manually. However, recent advancements in robotic technology [1,2] and the adoption of industrial manipulators have greatly improved the precision, speed and repeatability of sensor positioning and data acquisition, reducing the inspection times and reliance on certified inspectors [3]. Despite these advancements, data interpretation remains a manual, labour-intensive process prone to errors and misinterpretation, especially when dealing with large datasets where operator fatigue can become a critical factor.

Carbon Fibre-Reinforced Plastics (CFRPs) are composite materials consisting of a polymer matrix reinforced with carbon fibres, creating a material with an engineered structure manifesting a high tensile strength-to-weight ratio, and corrosion resistance [4]. As a result, CFRPs have become essential in industries such as aerospace, where they are utilised in constructing critical airframe components like fuselages, wings, and control surfaces, helping to reduce structural weight for improved fuel efficiency [5]. UT is the preferred method for bulk inspection due to its flexibility and safety, often used in the post-manufacturing inspection of CFRP components for quality control and generally the most widely adopted certification sign-off procedure for safety critical aerospace components [6,7]. UT operates on the principle of the piezoelectric effect, where an electric pulse excites a transducer to generate high-frequency ultrasonic waves. When a transducer is coupled to the test piece, the waves propagate into the material and interact with its internal structures. As the waves encounter scatterers and reflectors within the material, a portion of the waves' energy returns to the transducer. The received energies are converted back into voltage and recorded over time, forming an amplitude versus time graph known as an ultrasonic A-scan.

Phased Array Ultrasonic Testing (PAUT) builds upon the conventional UT by utilising an array configuration of ultrasonic transducers. This arrangement allows for precise electronic control of individual elements, where transmission/reception delays can be programmed to create electronic beamforming, linear scanning, and beam steering [8–10]. Usually, collected ultrasonic data is examined through different views. B-scan represents a two-dimensional cross-section (side) view of the inspected material, where brighter areas indicate higher signal amplitude of reflected/scattered wave. On the other hand, a C-scan is a top-down two-dimensional view containing maximum recorded amplitudes or Time-Of-Flight (TOF) information of the scanned component within a selected time gate. PAUT systems are prevalent in aerospace and energy industries due to their wider scanning coverage of the surface, and hence higher inspection speeds, advanced beamforming capabilities, and readiness for automated delivery [11,12].

Current industry NDE practices in the aerospace sector begin with automated robotic sensor delivery and data acquisition. This initial stage is followed by data preparation, which includes signal processing techniques such as frequency filtering, signal enveloping with the Hilbert transform [13] and signal gating. Next, NDE inspectors review segments of the C-scans, and if indications exceeding industry guidelines for allowable defect size or amplitude threshold are identified, the corresponding B-scans are further examined. Lastly, areas of interest are extracted for quality certification report creation. Automated robotic data acquisition for components like wing covers of midsize civil aircraft models typically takes around 40 min, with data analysis requiring a similar amount of time for pristine components. However, this step may be extended by an additional hour or more if artefacts and defects are detected. This additional time is allocated for further inspection of different views of the data, primarily individual B-scans around areas of interest, and the report generation process. The overall workflow is illustrated in Fig. 1.

Apart from defect detection, defect sizing is another critical step in the data analysis workflow. Current industrial guidelines for NDE inspection describe allowable defect sizes based on their type and location on the aircraft. For instance, in the case of delaminations, the largest allowable flaw area that would not be categorised as a defect range from 60 to 500 mm^2, depending on the specific location on the aircraft. Traditionally, defect sizing is achieved using the 6 dB drop method, where an operator manually moves the probe to find the maximum amplitude and then determines the defect boundaries by identifying points where the amplitude drops by 6 dB (i.e., to half of the maximum amplitude). This method allows for fine-tuned probe positioning, making it highly dependent on operator skill. A similar approach can be applied to automated PAUT testing. However, instead of manual movement, the PAUT array is manipulated using industrial robotics, significantly improving repeatability, precision, and scanning speed. Despite these advantages, the resolution for defect sizing is constrained by the fixed pitch between individual transducers and the predefined scan step.

Following detection and sizing, operators are tasked with categorising defects based on their physical properties, which are inferred from ultrasonic signal features. This classification step is critical in distinguishing between common defect types in CFRPs such as delaminations, porosities, and foreign object inclusions, each of which exhibits distinct patterns in ultrasonic data. This manual process is not only time-consuming and labour-intensive but also prone to inconsistencies as different operators may interpret the same dataset differently. The variability in human judgment introduces additional challenges in reproducibility and makes a fair assessment of performance difficult. The reliance on contextual judgment, global understanding of data, and external knowledge about the inspected components further highlights the complexity of the operator's role.

The above-mentioned tasks and workflow highlight the potential of automation in NDE data analysis, particularly in the aerospace industry, where large volumes of data are routinely handled. While data acquisition is predominantly automated, the subsequent stages of data analysis, defect identification, sizing, and classification, remain heavily reliant on NDE operators. In certain scenarios, basic automation tools can be used to analyse stable and well-defined signals. In Ref. [14], the authors introduce tools to assist with thickness measurements, detection of delaminations in areas with varying thickness, and evaluation of porosity content. These tools require human interaction to narrow down areas of interest and provide some input parameters, resulting in a reduction of analysis time by 70 %. Another approach is presented in Ref. [15], where data analysis is based on a multi-step algorithm. However, for complex signals heavily influenced by geometrical features of components, overlapping ultrasonic echoes, or external factors such as poor scan quality, the use of advanced solutions is needed [16].

Artificial Intelligence (AI) is an umbrella term that encompasses various algorithms and models designed to improve their performance with exposure to data. Machine learning is a subcategory of AI focused on developing advanced and complex models that utilise multiple layers of processing to transform input data. In recent years, there has been a notable increase in AI research addressing various NDE challenges, such as bridging the gap between simulated and real domains [17], automating image analysis [18], and enabling online path generation for robotic inspection [19]. Academic research in UT using AI can be

Fig. 1. Standard NDE workflow in the aerospace sector.

categorised by the different ultrasonic views used, namely A-scans, B-scans, and C-scans. Each view offers unique insights into the inspected material.

- **A-scan** view provides a simple representation of the signal's amplitude against the time of propagation. This view is useful for assessing the material thickness and identifying individual defects such as impact damage and larger delaminations [20]. However, A-scans are limited in spatial information and require expertise to interpret confidently. Most academic studies use A-scans as input for AI models, with a primary focus on the inspection of metal welds and steel samples. In Ref. [21], the authors have used Autoencoder (AE) architecture to denoise A-scans and improve their quality. The work presented in Ref. [22] utilised linear neural networks to determine fatigue life and tensile strength of spot welds from A-scan signals. The authors of [23] compared several classifiers on CFRP data, ultimately concluding that feature extraction using a Convolutional Neural Network (CNN) outperformed hand-crafted methods. In Ref. [24], the researchers employed a CNN paired with a gated recurrent unit to successfully classify manufactured debonding defects in braided CFRPs. Study detailed in Ref. [25] compared the performance of three models - CNN, Long Short-Term Memory (LSTM), and a combined CNN-LSTM model to precisely identify the defect depth in the CFRP sample, with the best model achieving an 8 % relative error compared to the ground truth.
- **B-scan** view offers a two-dimensional representation of the test material, providing spatial information and allowing for the visualisation of defects' location in relation to the material's geometry, making it generally easier to interpret than A-scans. Fewer academic publications focus on using B-scans as input to AI models, with most concentrating on metal materials. Object detection models for identifying defects in steel blocks, including EfficientDet, RetinaNet, and YOLO, were used in Ref. [26]. Building on this prior work, the authors in Ref. [27] improved defect detection results by simultaneously inputting three consecutive B-scans into the EfficientDet model to provide additional contextual information. In their latest research [28], the same authors further enhanced their approach by introducing a new model called DefectDet, specifically designed to address the challenges posed by extreme aspect ratios in B-scan views. A variational AE was employed in Ref. [29] to describe the distribution of pristine UT data and to discern defective B-scans based on the observed reconstruction errors. This research was further extended in Ref. [30], where the authors compared the anomaly detection performance of the GANomaly and PaDiM models and proposed a semi-supervised anomaly detection model called DifferNet.
- **C-scan** view presents indications associated with scatterers/reflectors in the form of amplitude or TOF across the entire scanned area, providing the spatial information in a 2D view. This view requires additional processing, including signal envelope analysis and careful gating. Although this view is often considered the easiest to interpret, its effectiveness may be compromised if defects are located close to the prominent geometrical features such as the front or back wall, as they could be erroneously gated out. Compared to A-scan and B-scan views, C-scans are the least used as inputs for AI models. In Ref. [31], the authors used TOF C-scans of aircraft components and modified YOLO family object detection models to identify defective areas. The work presented in Ref. [32] compared several AI models for the binary classification of various defects in reference CFRP samples, including U-Net, Transformer models, CNNs, and LSTMs, with U-Net achieving the best results. The authors of [33] explored the classification of fibre waviness in CFRP materials using U-Net and SegNet models, framing the problem as one of anomaly detection.

It has been demonstrated that AI models are capable of outperforming humans in certain tasks. The study detailed in Ref. [34] explored the capability of NDE inspectors to distinguish between real UT data and data created by generative AI models. The study concluded that artificial data is indistinguishable from real data, making it an ideal candidate for training future inspectors and for supplementation of training datasets for alternative AI models. In Ref. [35], the authors compared the defect detection performance of an AI model with that of three NDE operators. The results showed that human operators made a larger number of false calls, while the AI correctly identified all defects present in the data. This trend extends to other fields as well. In Ref. [36], the researchers demonstrated that an AI model designed for analysis and diagnosis of three-dimensional optical coherence tomography data matches or exceeds the accuracy of medical professionals with years of experience. Similarly, the researchers in Ref. [37] leveraged an ensemble of AI models to outperform human experts in medical diagnosis based on medical sonography. Despite the highlighted advancements in AI models, data analysis in industry remains predominantly manual, with limited adoption of new AI-based automation tools. Two key reasons for this are a lack of trust in the models, which includes concerns from both industry users and regulators, particularly in safety-critical processes [38], and the "black box" nature of AI, where the reasoning behind decisions is obscured. This lack of transparency leads to greater risks in evaluating safety-critical components, as inaccurate predictions from an automated system could result in unpredicted catastrophic in-service failures. Therefore, while these studies confirm the potential of incorporating new AI tools into NDE workflows, advancing to higher automation levels will depend on building trust in these systems.

Definitions of automation levels vary across fields and applications [39]. In the context of NDE, the authors of [16] propose a taxonomy for the entire NDE process, categorising it into Classical NDE (Level 0), Operator assistance (Level 1), Partial automation (Level 2), Operational automation (Level 3), and Full automation (Level 4). In recent years, there has been a notable shift towards adopting automated solutions in NDE workflows, leveraging advancements in robotics, AI, and other technologies, recognised as NDE 4.0 [16,40]. This transition aims to redefine the roles of human NDE operators, transitioning them to more supervisory positions where they oversee and address specific parts of the process, while automated systems manage the bulk of repetitive tasks. The overarching objective is to enhance efficiency while improving the precision and repeatability of the overall NDE workflow.

However, this evolution introduces several challenges. First, the increased complexity of automated systems can make troubleshooting and maintenance more difficult, as operators may need to develop new skills to manage these systems effectively. At the same time, the mental workload on staff is likely to increase [40]. Additionally, there is a risk of inappropriate reliance on automation, where tasks requiring human judgment are delegated to machines, potentially leading to errors or oversights. A study detailed in Ref. [38] assigned NDE operators detection and sizing tasks using automated tools and found significant levels of both disuse (operators disagreeing with the automation when it is correct) and misuse (operators agreeing with the automation when it is incorrect). To address this, the authors recommend incorporating discussions on the limitations of automation tools into the training of new personnel. Furthermore, by providing reasons behind potential automation failures, operators can develop a more informed and appropriate approach to using these tools, while also building trust through direct experience with the technology. In the context of automation, the term "human-in-the-loop" refers to systems where human operators remain actively involved in decision-making processes, while "out-of-the-loop" refers to systems where automation takes over tasks with no direct human involvement. Over-reliance on fully automated systems can result in out-of-the-loop performance degradation, where operators lose the ability to identify system errors and perform tasks manually. Studies, such as [41], have highlighted that operators relying on automation tools have diminished manual task performance

compared to those who perform tasks without automation. To address these issues, it is suggested that humans maintain a high level of control through periodic interventions, which can help minimise system failure rates [42].

Trust can be defined as subjective anticipation of future behaviour [43], often based on reported performance metrics on a subset of data used in the study. This approach shapes the human perception of trust, which is more effectively demonstrated through direct interaction with the model and observation of its decisions [44]. Some implementations leverage the human-in-the-loop method to enhance trust, where the human operator oversees and supervises decisions made by the AI, facilitating continuous improvement of the existing models. Such an approach was explored in Ref. [45] where humans collaborated with AI models to build trust and enhance accuracy. This was achieved by identifying anomalous instances of data, labelling them, and incorporating them into subsequent iterations of model development. Additionally, allowing the human operator to question and have control over AI predictions is another way to build trust in probabilistic models [46]. An alternative approach is to adopt explainable and interpretable AI [47]. However, these strategies are rarely explored in the field of NDE, with the most notable work being [48], where the authors used a novel dimensionality reduction method to strengthen the explainability of the AI model used for the sizing of defects from UT data.

While there is a clear need to increase automation in data analysis, and some progress has been made with traditional methods that offer significant time savings [14,15], guidelines for the practical implementation of AI tools in NDE are often lacking. Moreover, existing research on the adoption of AI methods for analysing UT data tends to focus on a single ultrasonic view. This approach does not accurately reflect how human inspectors conduct NDE, as they utilise multiple views to form conclusions about the inspected material. Relying on only one view can also overlook the strengths of other ultrasonic views, which may be better suited for inspecting varied locations, and detecting different types of defects or features. To address these gaps, this work focuses on.

- **Proposal and discussion of automation levels in data analysis**, ranging from operator assistance level (Level 1) to full automation (Level 4), with a focus on integration strategies to minimise the risk of critical system failures.
- **Development of a comprehensive PAUT data analysis workflow** utilising three distinct AI models that analyse B-scan views, C-scan views, and full 3D volumetric data in a coordinated manner.
- **Presenting a case study** involving an automated robotic inspection system for PAUT of CFRP materials used in the aerospace industry. This case study examines two reference industrial samples with complex geometry using an experimental setup that closely mimics industrial practices and employs industrial manipulators for accurate and precise measurements.

The rest of the paper is organised as follows: proposed levels of automation of data analysis are introduced in Section 2, Section 3 focuses on the materials and methods, Section 4 presents the results and discussion, and Section 5 concludes the work and outlines trajectories for future work.

2. Data analysis: levels of automation

2.1. Level 0: classical NDE

Taking inspiration from the automation levels defined for the entire NDE process defined in Ref. [16], Fig. 2 illustrates the proposed automation levels for data analysis. Data analysis at level 0 of automation corresponds to classical NDE, where the operator manually examines all data, performs preprocessing tasks, and makes decisions independently. This manual approach, although still widely used due to historical

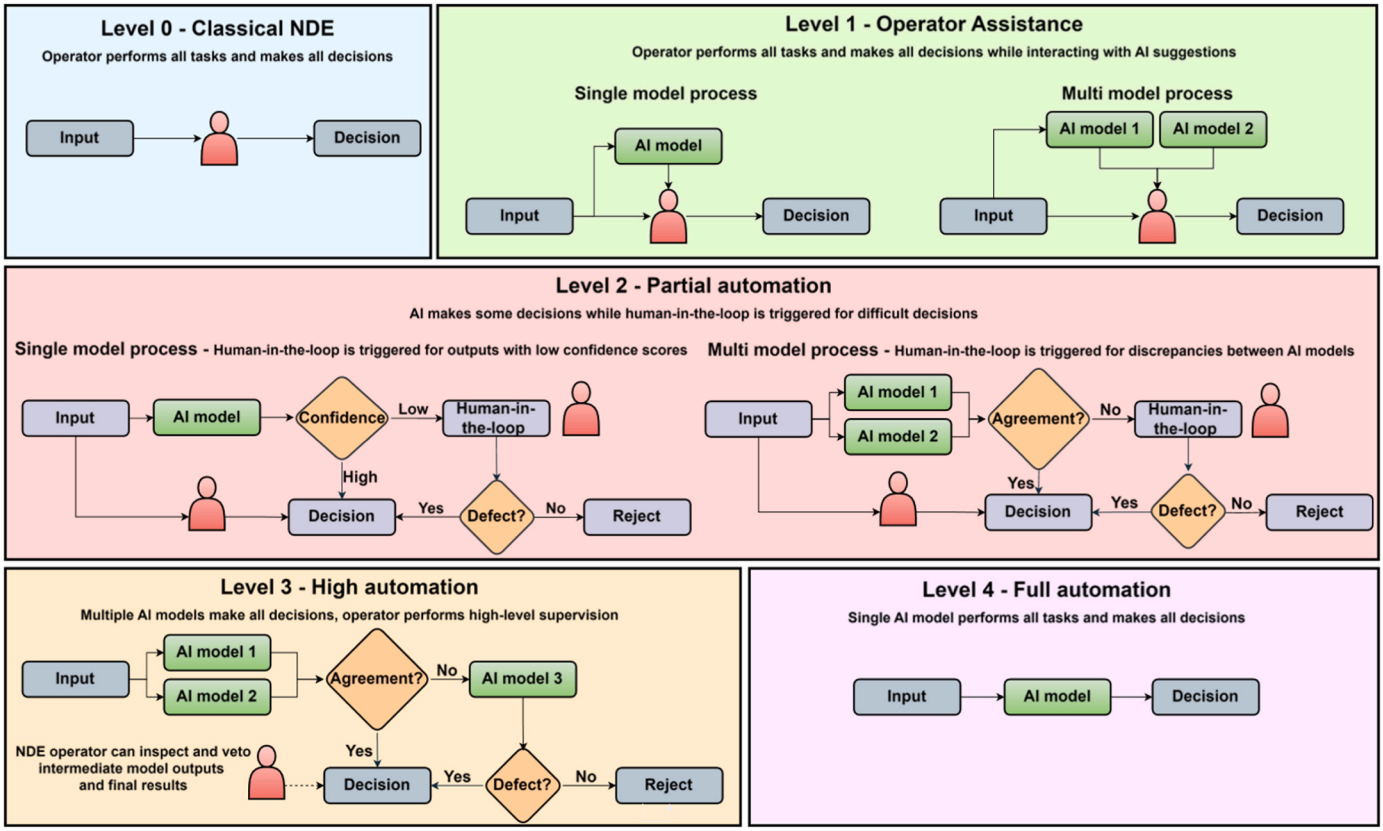

Fig. 2. Proposed data analysis workflows for different levels of automation.

industrial approach in training operators and reliance on individual decision making, relies heavily on the operator's NDE expertise and judgment. While offering high traceability and explainability, it also results in longer data analysis times, higher operator workload, and increased likelihood of human-induced errors, particularly during prolonged repetitive tasks [40].

2.2. Level 1: operator assistance

At Level 1, operator-assisted data analysis, the human operator retains responsibility for all decisions and tasks. AI models assist by providing suggestions and highlighting areas of interest, but the final decision regarding the presence/absence of defects (sentencing) remains with the operator. However, the risk of failure at this level of automation is higher than at Level 0, primarily due to the potential for inappropriate reliance on automation, which could lead to misuse or disuse of the AI tools. Therefore, this level requires moderate trust in the AI models, which are expected to generate suggestions while focusing on minimising false negative (failing to detect an actual defect) calls to accelerate the analysis process. When used correctly, operator-assisted automation is ideal for gaining insights into scenarios where AI models may underperform without compromising the quality of the final NDE inspection. Additionally, this approach allows for continuous improvement by using those findings in future model re-training. Lastly, allowing operators to interact with the models, fine-tune inference parameters, and observe outputs during deployment could help build trust over time, as suggested in Ref. [44]. However, it is important to note that, as AI technologies are not yet widely implemented in the industry, NDE operators have not been trained in refining or adjusting AI tools. Therefore, additional training would be necessary for operators to confidently undertake this task.

2.3. Level 2: partial automation

Partial automation of NDE data analysis at Level 2 relies on a combined system of single- or multi model processing with human-in-the-loop decision-making. For single model setup, predictions with confidence scores above a set threshold are accepted automatically, while lower-confidence predictions are passed to a human-in-the-loop mechanism for further review. On the other hand, multi model configuration involves two AI models collaborating to identify areas with potential defect indications, automatically accepting them if their decisions coincide, and activating human-in-the-loop decision-making to resolve any disagreements. This approach greatly accelerates data analysis by focusing human intervention solely on resolving model discrepancies rather than manually processing all data.

The prerequisite for this level of automation is a high trust in the models to identify all defective areas while tolerating some false positives. False positives are managed through two mechanisms: first, by cross-verifying outputs between two detection models, which are unlikely to produce identical false positives, and second, by engaging human-in-the-loop decisions when models disagree. Overall, Level 2 of automation is characterised by faster data analysis and reduced human workload, albeit at the expense of higher system complexity and an elevated risk of failure. To prevent human-out-of-the-loop performance degradation, operators retain the ability to intervene and take control at any time. They can audit AI decisions and examine intermediate outputs from each stage, thereby improving both explainability and traceability.

2.4. Level 3: high automation

Level 3 automation operates as a multi model system, where a third, higher-precision AI model resolves disagreements between the initial two AI models. In this work, the two initial models are selected based on a logic of mirroring the manual approach taken by human operators, who typically examine C-scan data first to identify defects and then use B-scan data for further investigation. Therefore, the two models work independently on C-scan and B-scan data, and with their rapid inference offer a balanced combination of efficiency and accuracy for defect detection. The third model, which operates on full 3D volumetric data, offers the highest precision and is reserved for the final verification of areas where the first two models disagree. However, it is the slowest of the three and scales the least efficiently with increases in data size. Instead, it is selectively applied to specific sections, replacing the human-in-the-loop mechanism from Level 2. At this level, used AI models must be scrutinised and fine-tuned, aiming to achieve optimal accuracy with no tolerance for false negatives. Human operators, while removed from direct involvement, transition to a supervisory role, retaining the ability to intervene, monitor, and override AI decisions as necessary. This configuration delivers many advantages of an ideal automated system, albeit with slightly slower analysis and increased computational power required to run multiple AI models in parallel.

2.5. Level 4: full automation

Level 4 automation represents an idealised long-term goal where an AI model surpasses human capabilities in both speed and accuracy. In this setup, a single end-to-end model is responsible for all decision-making, eliminating the need for human NDE operators to inspect the data. While this approach would offer the fastest analysis, it comes with the highest risks and requires very high trust in the AI model, which can only be achieved through rigorous testing and parameter tuning. This level also represents an extreme case of automation, where human out-of-the-loop performance issues might arise. Most academic works that report evaluation metrics for developed AI models can be seen as examples of Level 4 automation.

It is important to note that different inspection scenarios may benefit from different combinations of AI models depending on requirements such as inference speed, precision, and explainability. As automation levels increase, several key system characteristics change. Higher automation levels lead to faster analysis speeds, with significant reductions in human workload. However, this comes at the cost of increased risks and system complexity. Lastly, trust in the AI system becomes crucial at higher levels of automation. Table 1 provides an overview of these system characteristics across different data analysis automation levels.

The lower risk associated with human operator performance stems from their ability to demonstrate inspection competency through rigorous training and testing. This acquired expertise is expected to generalise to out-of-distribution cases, as it is based on fundamental principles rather than solely on pattern recognition. In contrast, AI-based approaches often struggle with out-of-distribution scenarios, leading to higher inspection risks. However, the risk level for human operators is not fixed and varies significantly depending on individual

Table 1
System characteristics for different automation levels of data analysis.

System Characteristic	Automation Level				
	Level 0	Level 1	Level 2	Level 3	Level 4
Risk	Operator dependent	Low	Medium	High	Very High
AI Trust	None	Low	Medium	High	Very High
System complexity	Very low	Low	Medium	High	Very High
Human workload	Very High	Very High	Medium	Low	None
AI workload	None	Low	Medium	High	Very High
Analysis speed	Very slow	Slow	Medium	Fast	Very Fast

skill and experience.

3. Materials and methods

3.1. Experimental setup

For data acquisition, an Olympus/Evident RollerFORM-5L64 [49] phased array probe, which is a roller probe product suited for automation of PAUT, paired with Peak NDT Ltd. MicroPulse 6 controller [50] was used. This phased array with a central frequency of 5 MHz comprises 64 individual ultrasonic elements arranged linearly with a pitch of 0.8 mm and an elevation of 6.4 mm, positioned inside a deformable tyre filled with a liquid creating a 25 mm stand-off between the array and the sample's surface. The liquid was selected to closely match the acoustic impedance of the tyre to facilitate the propagation of ultrasonic waves into the sample.

The ultrasonic controller features 128 transmission and reception channels, allowing for the customisation of focal laws. For this study, a linear scanning mode with a sub-aperture of 4 elements was employed. This resulted in an active aperture of 48.8 mm and the recording of 61 A-scans in each electronic sweep of the array. An excitation voltage of 80 V and pulse width of 100 ns were used, in conjunction with a digital 6 MHz lowpass filter and a sampling rate of 100 MHz. An overall gain of 22.5 dB was applied upon reception of the signal, in addition to Time Variable Gain (TVG) added during post-processing. The use of TVG enhances the signal amplitudes in the later stages of the ultrasonic propagation, compensating for the highly attenuative nature of the inspected CFRP material, as is set to 1.5 dB/mm.

The PAUT sensor was delivered to scan the component through an automated platform built around an industrial manipulator KUKA KR90 R3100 extra HA. To achieve a consistent coupling quality of the PAUT roller probe to the component's surface during the scan, a bespoke robotic control platform was used to adaptively change the robot pose based on real-time feedback from a Schunk GmbH & Co. FTN-GAMMA-IP65 SI-130-10 force torque sensor [51] mounted on the robot [52]. Air pockets between the probe's tyre and the sample's surface act as strong reflectors on the UT signals, diminishing the sample's volumetric signal quality, therefore sufficient coupling had to be achieved and sustained during the scan. To this end, water was sprayed over the sample surface, creating a thin film between the material and the PAUT tyre.

3.2. Reference samples

This study focuses on the analysis of two CFRP samples manufactured by Spirit AeroSystems, UK according to Bombardier aerospace process specification standard. To partly imitate defects occurring in the manufacturing process, a range of Polytetrafluoroethylene (PTFE) and other polymer inserts were embedded [53]. The first sample (Sample A) was a stepped specimen with dimensions 780.0 × 200.0 mm and thicknesses ranging from 7.5 to 13.5 mm. Square inserts, each measuring 6.0 × 6.0 mm, were embedded in the sample at different depths and locations, resulting in a total of 24 defects. The inserts were positioned both near the edges and at the centre of the sample, with depths ranging from subsurface levels to near the back wall, as detailed in.

Table 2. An amplitude C-scan, model of the sample A, and PAUT roller probe dimensions are presented in Fig. 3.

The second sample (Sample B) was composed of a flat panel skin surface co-cured with three stringer sections. The sample contained 12 PTFE inserts, with 6 located immediately beneath the surface and 6 beneath the stringer sections, as detailed in Table 3. The sizes of the inserts were 20.0 × 10.0 mm, 10.0 × 5.0 mm, and 5.0 × 5.0 mm. An amplitude C-scan and a model of the sample B are illustrated in Fig. 4.

3.3. Data stream handling

Training and development of AI models was performed on a high-performance desktop Windows 11 PC. This system was equipped with an Nvidia RTX 3090 Ti Graphics Processing Unit (GPU), 128 GB of Random Access Memory (RAM), and two Intel® Xeon® Gold 6428 2.50 GHz Central Processing Unit (CPUs), while PyTorch [54] framework was used for model development. The data processing, data capture, synchronisation between individual hardware elements, and model inferences were performed on a Windows 11 Dell Precision 5570 laptop equipped with Intel i9-12900H 2.50 GHz CPU, 64 GB of RAM, and NVIDIA RTX A2000 8 GB GPU. Robotic control was executed with JAVA code wrapped in Python syntax, while UT and AI processes were performed in Python 3.8. Data acquisition and processing were split into two Python nodes.

The acquisition/communication node first sets up a Transmission Control Protocol/Internet Protocol (TCP/IP) connection to the UT equipment and listens for the User Datagram Protocol (UDP) connection established by the KUKA robotic controller. Once all connections are active, another TCP/IP connection to the processing node is initiated to maximise the utilisation of available computing resources. In Python, the global interpreter lock restricts true multiprocessing and parallelism within a single interpreter process, therefore running two separate scripts concurrently allows the scripts to utilise different CPU cores effectively. Alternative programming languages such as C++ offer more straightforward solutions for parallelism but would result in more complex code and make implementation of the developed AI models more difficult. Another potential solution includes using Robotic Operating System (ROS) framework which is optimised for real-time applications.

Upon establishing the connection to the processing node, the robotic controller begins monitoring and broadcasting its' positions. The communication node continuously checks the Euclidean distance between subsequent position updates, triggering the UT data capture command if it surpasses the predetermined distance threshold of 0.8 mm (*i.e.* scanning step used in this study). The threshold aligns with the pitch of the used PAUT assembly, ensuring a square aspect ratio in the final data representation. Upon receiving the data, the robotic positions and

Table 2
Details for sample A.

Defect	Between plies (start/end)	~Depth (mm)	Sample thickness (mm)	Defect	Between plies (start/end)	~Depth (mm)	Sample thickness (mm)
1	2/3	0.65	13.50	13	18/19	4.80	9.59
2	2/3	0.65	13.50	14	18/19	4.80	9.59
3	2/3	0.65	11.70	15	14/15	3.76	7.46
4	2/3	0.65	11.70	16	14/15	3.76	7.46
5	2/3	0.65	9.59	17	50/51	13.11	13.50
6	2/3	0.65	9.59	18	50/51	13.11	13.50
7	2/3	0.65	7.46	19	42/43	11.03	11.70
8	2/3	0.65	7.46	20	42/43	11.03	11.70
9	26/27	6.88	13.50	21	34/35	8.96	9.59
10	26/27	6.88	13.50	22	34/35	8.96	9.59
11	22/23	5.84	11.70	23	26/27	6.88	7.46
12	22/23	5.84	11.70	24	26/27	6.88	7.46

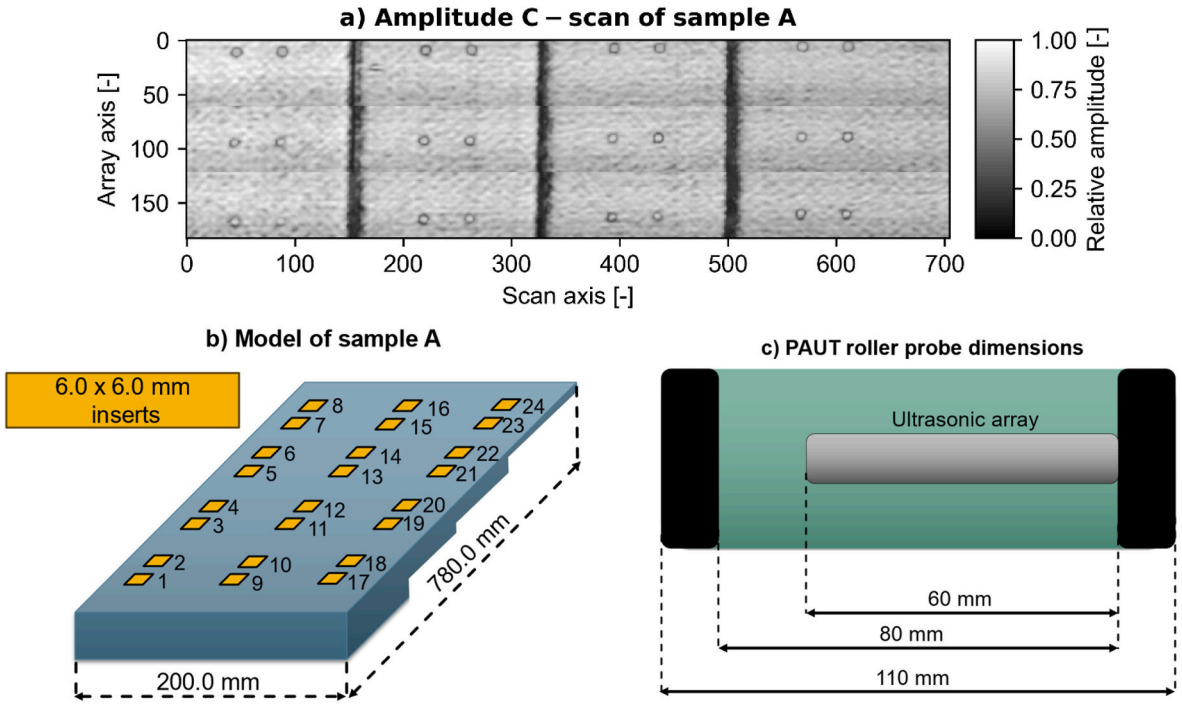

Fig. 3. A) Amplitude C-scan of sample A, b) Model of sample A, c) PAUT probe dimensions.

Table 3
Details of defects in sample B.

Defect	Insert size (mm x mm)	Between plies (start/end)	~Depth (mm)	Sample thickness (mm)	Defect	Insert size (mm x mm)	Between plies (start/end)	~Depth (mm)	Sample thickness (mm)
1	5.0 × 5.0	2/3	0.9	7.8	7	5.0 × 5.0	18/19	6.8	12.5
2	10.0 × 10.0	2/3	0.9	7.8	8	10.0 × 10.0	18/19	6.8	12.5
3	20.0 × 10.0	2/3	0.9	7.8	9	20.0 × 10.0	18/19	6.8	12.5
4	5.0 × 5.0	2/3	0.9	7.8	10	5.0 × 5.0	18/19	6.8	12.5
5	10.0 × 10.0	2/3	0.9	7.8	11	10.0 × 10.0	18/19	6.8	12.5
6	20.0 × 10.0	2/3	0.9	7.8	12	20.0 × 10.0	18/19	6.8	12.5

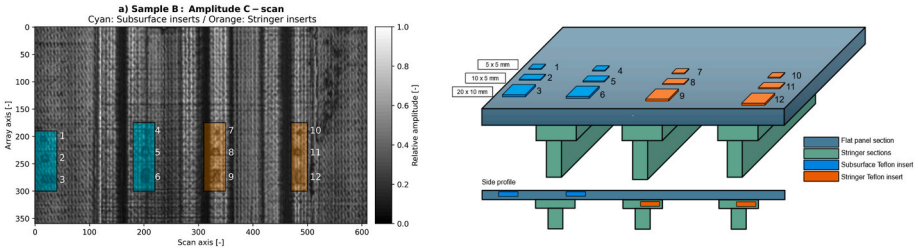

Fig. 4. Amplitude C-scan and a schematic of Sample B demonstrating the actual and estimated location of defects, respectively.

UT readings are correlated and transmitted immediately to the processing node, repeating the process until the scan is completed. The processing node, upon receiving the data, performs basic data manipulation, including reshaping, normalisation, data padding, TVG, and Hilbert transform, before feeding the data into AI models described in Section 3.4. A block diagram illustrating the ultrasonic and robotic setup, alongside the communication and processing nodes, is illustrated in Fig. 5.

Additionally, it is important to address the limitations and set reasonable expectations for the positional triggering setup. The current configuration, with a UDP connection between the KUKA controller and laptop, provides a positional update rate of 250 Hz. This update rate may present challenges at higher scanning speeds, potentially resulting in positional overshooting for data capture triggers. In the conducted experiments, scanning speeds of up to 30 mm/s were tested and deemed satisfactory.

Another critical aspect to consider is the resolution of ultrasonic scans. In the aerospace sector, the primary objective of NDE is to detect defects classified as critical based on their size and location on the structure. Quality control documents from Spirit AeroSystems indicate that delaminations ranging from 60 to 500 mm^2 may be allowed, depending on their position within the structure. When converted to equivalent circular defects, these areas correspond to defect diameters ranging from 8.8 to 25 mm. Given this context, acquiring data at intervals of 0.8 mm ensures at least five frames per defect are captured.

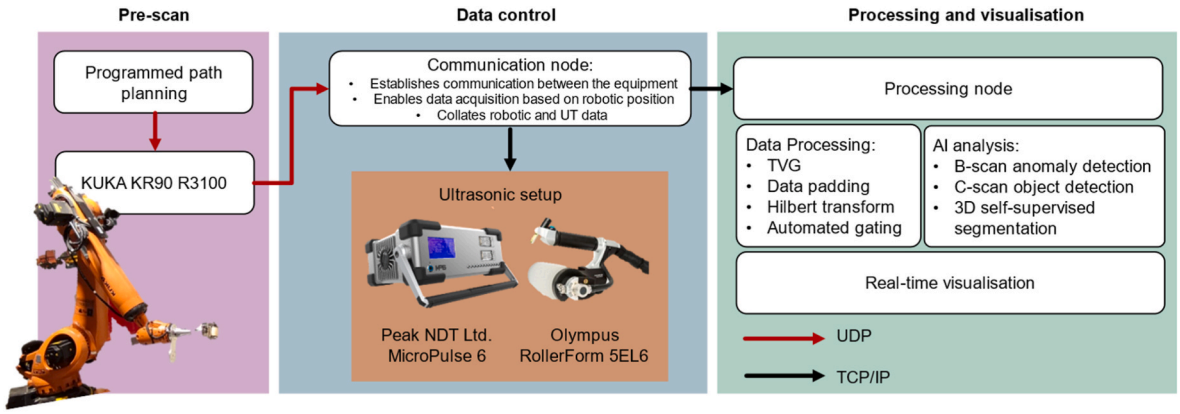

Fig. 5. Flowchart of the experimental setup, integration of PAUT and robot, and data flow.

3.4. Artificial Intelligence models

3.4.1. Anomaly autoencoder model

The first AI model used in this work is detailed in Ref. [55]. The framework comprises a two-step process: an automated gating method and an anomaly detector based on the AE structure. While the surface echo is removed automatically due to a constant known offset of the inspected sample from the ultrasonic array, determined by the roller probe's outer diameter, the automated gating method analyses volumetric data to identify and exclude the back-wall echo. This approach operates without needing prior knowledge of the material's thickness or geometry and is achieved using a peak-finding algorithm from the SciPy Python library [56] applied to Hilbert-transformed volumetric data. The algorithm applies a threshold of 0.25 for normalised amplitudes, a value chosen within the range of 0–1 and calibrated using an additional flat CFRP pristine sample, and a minimum distance of 5 time samples between identified peaks to filter out minor peaks thus reducing data dimensionality and processing times. The identified peaks are subsequently processed using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, an unsupervised clustering algorithm [57], which groups peaks into distinct clusters based on two parameters: *eps* (the maximum allowable distance between peaks in a cluster) and *min_number_of_peaks (*the minimum count of peaks required to form a cluster). The *eps* value is set to 7, following the values from the original publication. The value for *min_number_of_peaks* is manually defined to ensure that defects up to 20.0 × 10.0 mm in size are not excluded. It is set to 250 to capture significant defects while excluding smaller, irrelevant clusters. This threshold can also be set automatically by observing the size of the captured data and adjusting the value based on the number of expected changes in the material's thickness. Finally, identified clusters, representing the back wall echo, are removed from the data.

Afterwards, the gated data is passed through the AE, which consists of encoder and decoder components, as illustrated in Fig. 6. The encoder is composed of a series of convolutional layers that compress input B-scans into a feature representation, while the decoder performs the inverse operation to reconstruct the input. Since the training dataset exclusively consists of pristine data, the model learns the expected structural noise patterns and is able to reconstruct these without issue. However, when a B-scan containing defects is introduced, the model struggles to accurately reconstruct these anomalies. The discrepancy between the input and output is quantified using Mean Squared Error (MSE), where a higher MSE indicates the presence of potential defects. To differentiate pristine from defective B-scans, an anomaly threshold is applied to the observed MSE errors. A single threshold is applied across all automation levels, set as the median value of all observed MSE errors, increased by 50 % of the median value. This approach is based on the expectation that most B-scans in the scanned sample are pristine. As a result, the median value of MSE will effectively represent the typical value for pristine B-scans, while the additional offset helps capture only significant deviations, accounting for smaller variations. This method is consistent across all levels of automation, prioritising the safety considerations and detection of defects while accepting a higher rate of false positives. It is worth noting that this threshold can be adjusted based on the specific application scenario, and could also be defined using other statistical methods, such as standard deviations.

This unsupervised learning approach demonstrated successful defect detection for defects larger than 4.0 mm in diameter. The model's lightweight architecture enables efficient integration without significant computational demands, as inference on an NVIDIA 3080 Ti GPU-accelerated machine can batch-process approximately 2000 B-scans in 1.26 ± 0.09 s. The limitation of this model lies in the failure to detect smaller defects (approximately 3.0 mm in diameter) and those located within 0.5 mm of the material's back wall. Additionally, this approach

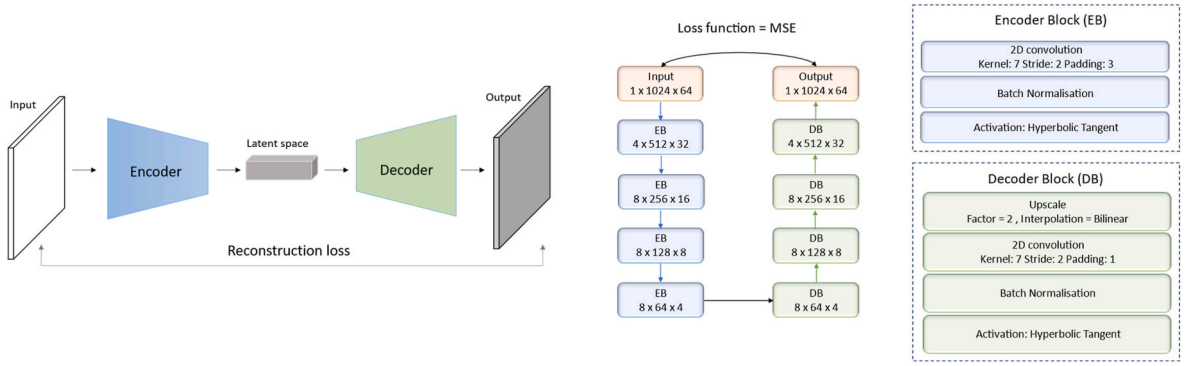

Fig. 6. Autoencoder architecture and specifications.

provides only a rough estimation of defect locations, as each B-scan is classified as either anomalous or not, without offering precise information about the exact position of the defect within the B-scan view. Lastly, removing the back wall signal eliminates useful information that NDE operators often rely on during analysis, particularly when assessing defects such as porosities, which are typically identified by a reduction in back wall amplitude.

3.4.2. Object detection model

The second model utilised in this work was a Faster Region-Convolutional Neural Network (R-CNN) object detection model detailed in Ref. [58]. Originally introduced in Ref. [59], Faster R-CNN is used as an end-to-end method for detecting various objects within image data. Its architecture consists of a convolutional feature extractor, a region proposal network that identifies regions of interest, and a classifier for object identification, as illustrated in Fig. 7.

The supervised training of Faster R-CNN required simulated datasets, which involved substantial computation time, manually crafted ground truth labels, and domain-specific augmentations to produce a high-performing model [17]. Despite its complex architecture, demanding training process, and higher computational requirements during inference, Faster R-CNN remains suitable for real-time deployment, achieving an inference speed of under 50 ms per C-scan on an NVIDIA 3080 Ti GPU-accelerated system. During deployment, Faster R-CNN requires a confidence threshold to filter generated predictions (value between 0 and 1). Following the same logic for setting anomaly thresholds for the AE model, confidence thresholds are set at 0.001 for all automation levels. The inference process begins with the generation of amplitude C-scans using the automated gating method described in section 3.4.1. These C-scans are then fed into the Faster R-CNN model, which outputs bounding boxes that highlight the defects within the inspected material. Compared to the AE model, the Faster R-CNN offers superior detection performance and provides the ability to precisely locate defects in the inspection plane.

The primary drawback of this model is its “black box" nature, where the reasoning behind inference results is obscured. In industry sectors requiring clear, interpretable outputs, this is a disadvantage, as it limits transparency in the decision-making process. Furthermore, since the model was trained on 64 × 64 resolution images, it can struggle when processing input images with significantly different aspect ratios or sizes. To overcome this challenge, a workaround involves applying the model on smaller sections of the scans and then collating the results.

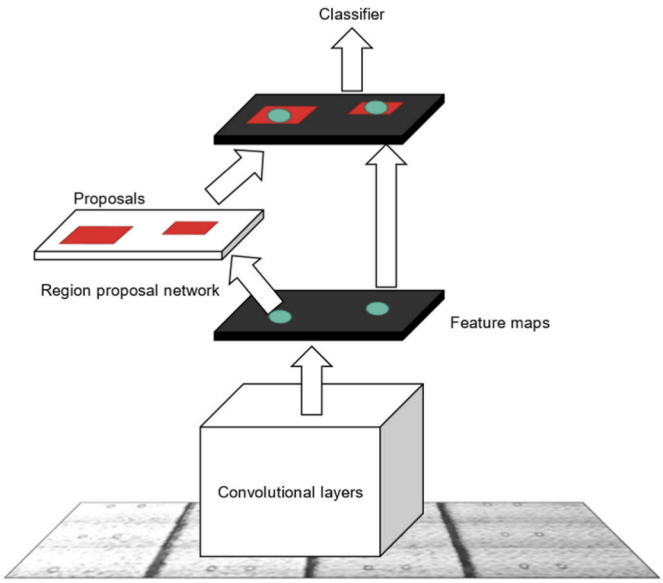

Fig. 7. Faster R-CNN architecture (adapted from Ref. [59]).

Although this slightly complicates the deployment of the code, it allows for efficient inference and reliable defect detection. Lastly, initial testing revealed that Faster R-CNN occasionally struggled to detect defects as small as 3.0 mm, especially when their amplitude was very low.

3.4.3. Self-supervised model

The third model was a 3D Ultrasonic Self-Supervised Segmentation (3-DUSSS) model designed to process full 3D volumetric data, as presented in Ref. [60]. This lightweight model operates by pre-training on pristine 1-D scan series through the component, where the model attempts to predict the likely distribution for the next value in the sequence. During inference, these distributions can be thresholded against the measured experimental values and used to indicate anomalous voxels. The model utilises a sliding window approach, whereby if a point is considered defect free it is added to the series to ground the model in relation to experimental data. If the point is considered defective, the model uses the mean of the predicted distribution as a best proxy for the expected defect-free datapoint and flags the voxel as defective. Similar to the AE model, training was performed on pristine data, allowing the 3-DUSSS model to learn the amplitude responses specific to carbon fibre structures. The training dataset included both front and back walls, which minimises the impact of poor gating which could lead to removal of defect signatures. During inference, the model requires two parameters: the allowable false call rate and an area threshold. The allowable false call rate defines the maximum deviation a voxel can have from the predicted distribution before being considered defective (in this work, this was set to 0.999999). The area threshold filters out smaller voxel groups to minimise false positive calls, with this threshold set to 10 in the current work.

The developed model excels in localisation, depth estimation, and sizing of defects, effectively detecting flaws as small as 3.0 mm in diameter. However, when processing large datasets, this method encounters challenges due to the computational demand of handling the entire scan volume, requiring a powerful GPU with significant memory capacity (the original study employed a setup with three NVIDIA GeForce RTX 3090 GPUs). Although the model itself is lightweight, the volume of data for processing is substantially higher than that of individual B- or C-scan views, making GPU memory a critical factor and creating a bottleneck in data loading onto the GPU. Even after downsampling data by a factor of 10 in the time domain, deployment on less powerful hardware, like the single GPU configuration used in this study, leads to processing times in the range of several minutes, far slower than the few seconds needed by AE and Faster-RCNN. Furthermore, the scans in this study are relatively small compared to those typical in industrial settings for large components, where AE and Faster-RCNN would likely scale better, as 3-DUSSS must process the entire dataset, while other methods operate on compressed 2D views. Additionally, 3-DUSSS faces challenges when encountering variations in thickness, making it more suitable for deployment along scan directions where thickness changes are minimal. However, despite being slower at inference than other methods, 3-DUSSS’s capability to generate a complete 3D segmentation map provides a comprehensive visualisation of the ultrasonic scan. This feature not only improves the interpretability of scan results but also allows for the creation of digital twins for reporting, adding practical value to the inspection process.

4. Results and discussion

4.1. Level 1 – operator assistance

In the Level 1 Operator Assistance level of data analysis, inference parameters for both the Faster R-CNN and AE models are configured to minimise the risk of false negatives. While the ideal performance of an NDE operator or automated system would result in zero false negatives and false positives, achieving this is challenging. In the context of NDE for safety-critical components, the emphasis is heavily on minimising

false negatives. Missing critical defects can have severe consequences, while false positives, though not directly threatening to material safety, may lead to higher costs if unnecessary rework is done, components are scrapped, or extended data analysis is conducted by operators to verify whether an indication is a true positive. For Faster-RCNN, the confidence threshold determines the number of defects identified: a higher threshold results in fewer, but more confident detections, reducing false positives but potentially missing smaller or more subtle defects. On the other hand, a lower confidence threshold increases the number of detections for smaller defects and fainter indications, though this often leads to more false positives. At this automation level, where the final decision rests with the operator and all data is expected to be reviewed, the preference is typically for a lower confidence threshold (i.e., 0.001). This setting helps minimise false negatives while relying on operators to review and filter out false positives, ensuring that potential defects are flagged for further inspection and prioritising safety by reducing the risk of overlooked critical defects.

For the AE model, inference involves setting a threshold for anomaly detection based on observed MSE. A higher threshold flags only severe discrepancies from the MSE associated with pristine B-scans (i.e., significant defects), thus reducing false positives but potentially missing minor defects. A lower threshold, on the other hand, captures a larger number of indications, including minor deviations that may represent pristine B-scans, increasing the risk of false positives. Following a similar approach to the Faster R-CNN model, the AE model at this level of automation is configured to prioritise safety by applying a lower

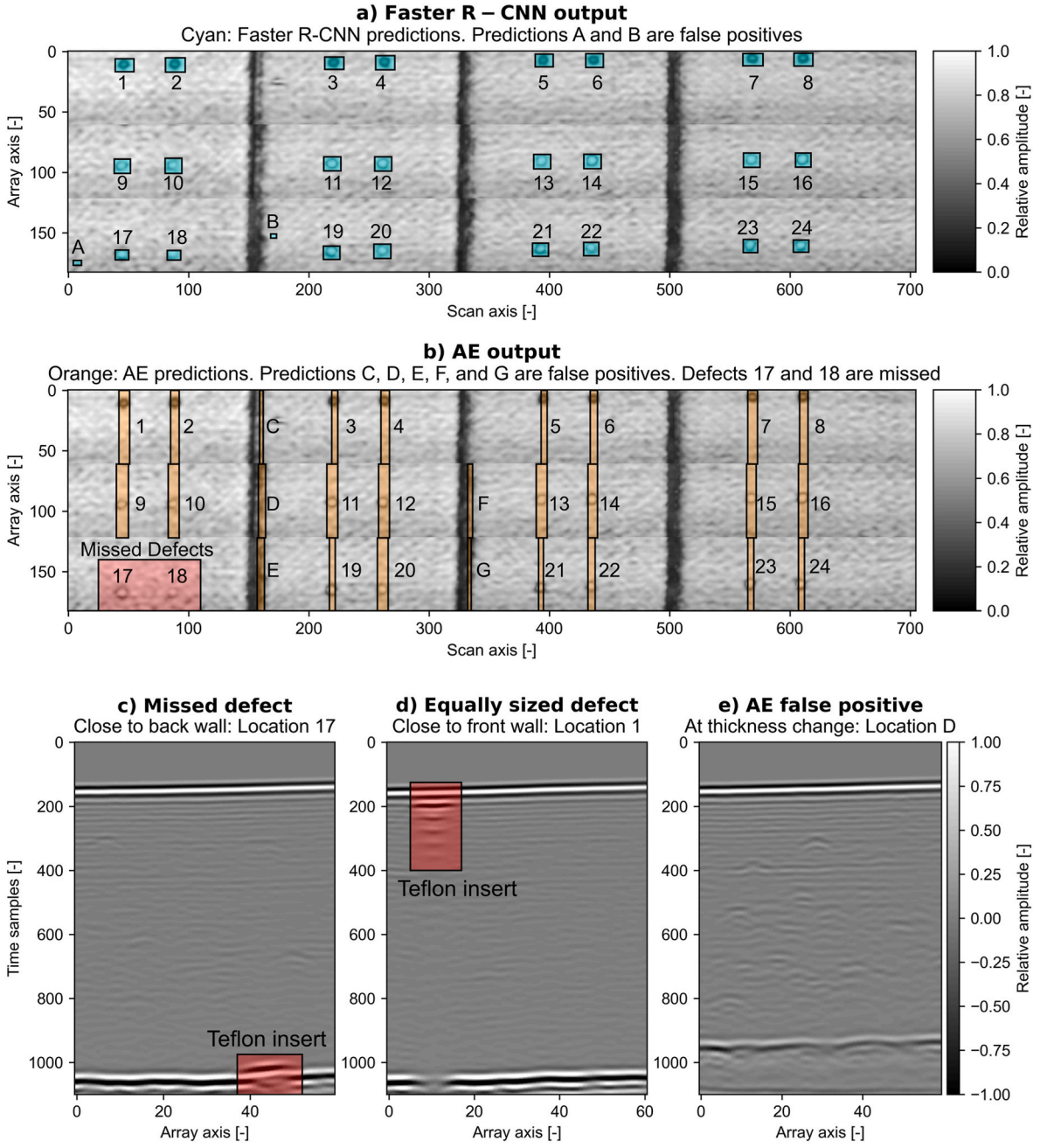

Fig. 8. A) Output of the Faster R-CNN model, and b) Output of the AE model on C-scan view of the sample A (cyan/orange bounding boxes); c) B-scan frame containing a missed defect indication close to back wall; d) Equally sized defect close to front wall with ultrasound reverberations aiding the defect detection; e) AE false positive resulting from minor indications received from thickness transition at the location of sample geometrical steps. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

anomaly detection threshold (i.e., median MSE plus 50 %).

The detection results for Sample A from both the Faster R-CNN and AE models are illustrated in Fig. 8 a) and b). Faster R-CNN was able to capture all defects while producing two false positives. However, the AE model missed two defects, located at the thickest section of the sample near the back wall. A B-scan from that position is shown in Fig. 8 c), where it can be observed why defects at such locations complicate detection. The indication is nearly fused with the back wall echo; Therefore, even with an optimal gating approach, a part of the defective signal would also be removed. This presents a challenge, as the defect appears very small on the B-scan level, and the C-scan amplitude response is also considerably weaker compared to other defects.

As a comparison, Fig. 8 d) shows a defect of equal size located immediately after the surface echo. While the main defect echo is again partially merged with the surface echo, as seen in the previous scenario, the several recorded ultrasonic repeats of the interface with a defect make detection easier. Therefore, even imperfect gating would leave strong reflections in the data, resulting in easier detections from both AE and Faster R-CNN. Lastly, Fig. 8 e) presents a B-scan showing a false positive indication produced by the AE model. In this instance, a change in sample thickness results in many higher amplitude reflections caused by the interaction between the ultrasonic beam and sharp transition in sample geometry. Therefore, while the area captured in the B-scan frame is pristine, these minor indications cause a substantial deviation from the median MSE observed in the rest of the scan, resulting in a false positive flag.

In the presented examples, models are prone to generating false positive or false negative indications when calibrated with lower confidence and anomaly thresholds. Unfortunately, this approach yields results unsuitable for higher automation levels, especially due to the risks associated with missing defects, which could compromise the structural integrity of the final product if left unchecked. While false positives degrade the inference performance, they do not pose direct safety risks and can be addressed by NDE operators, albeit at the cost of additional analysis time. Nevertheless, the primary aim of this automation level is to assist with the analysis by providing informed suggestions on areas of interest, with the final decision remaining with the NDE operator who reviews all data. While adjusting model inference parameters could potentially lead to the successful detection of all defects by both models, changing these values on per sample basis is not feasible in the industrial system deployment.

The reasoning behind choosing the AE and Faster R-CNN models as the primary models for this application is their fast inference times, making them suitable for deployment on less powerful hardware. Only the inference times of the models are reported in this work. Faster R-CNN processing for Sample A takes 0.22 ± 0.06 s, while the AE model produces results in 2.28 ± 0.12 s. Specifically, the inference time for the AE is 1.56 ± 0.01 s, with an additional 0.73 ± 0.025 s required for padding inputs to match the AE's convolutional structure. On the other hand, running the 3-DUSSS model on sample A takes 221.34 ± 1.41 s. The results of the 3-DUSSS model are overlayed over a C-scan of the sample and presented in Fig. 9.

The detection results for Sample B are illustrated in Fig. 10 a) and b). While Faster R-CNN successfully identified all defects with two false positives, the AE model failed to detect a 5.0 × 5.0 mm defect in the stringer section. Upon further inspection, this defect is partially visible in the scans but was not captured in its entirety due to an insufficient overlap between adjacent ultrasonic passes. As a result of this scanning error, the defect appears smaller than its true size, reducing its amplitude response, and preventing it from meeting the detection threshold for the AE model. Although reducing the anomaly threshold further might enable detection of this defect, it would also result in an excessive number of false positives across the scan. This defect is shown in Fig. 10 c).

Additionally, while AE successfully identifies defects near the front wall, not all B-scans containing defects are flagged. Since defects typically span several B-scan slices, the MSE error varies across these slices, leading to some B-scans being correctly classified as anomalous while others are not. This approach still serves its purpose, as it provides the operator with a highlighted area of interest, which is valuable for guiding further inspection (although achieving complete detection would be ideal). An example of a partially captured defect is in Fig. 10 d). This example highlights the advantage of Faster R-CNN, which leverages the spatial context across the C-scan view, rather than relying solely on individual B-scan slices.

Inference for FasterRCNN took 0.55 ± 0.08 s, while AE produced results in 2.34 ± 0.11 s, with additional time for padding resulting in 0.67 ± 0.01 s 3-DUSSS model for this larger sample runs in 379.98 ± 1.21 s, which underscores the challenges in the scaling of inference time. In contrast, manual NDE inspection is typically reported to take significantly longer. For example, for a sample approximately double the size, the data interrogation is typically completed in 40 min by an operator, although this time is extended by an hour or more when defects are present as a closer examination and sizing of defective areas is required. While direct measurements for human analysis of the specific samples discussed in this work are not available, these figures highlight the time-saving potential of the proposed AI-based methods, which operate on the scale of seconds and minutes compared to tens of minutes or hours for manual inspection. The results of the 3-DUSSS model are overlaid on a C-scan and presented in Fig. 11, showing that all defects were successfully detected, with five false positives.

4.2. Level 2 – partial automation

Level 2 of automation combines and compares the outputs of models, adding a layer of validation to AI predictions. In sample A, Faster R-CNN and AE agreed on 22 out of 24 defects, as shown in Fig. 12. This agreement enhances trust in the system, as these areas are flagged by two independently trained AI models, each trained on distinct data and ultrasonic views. Meanwhile, the nine areas of disagreement were flagged for human review, streamlining the analysis process. Rather than examining the entire dataset, the operator can now focus on these

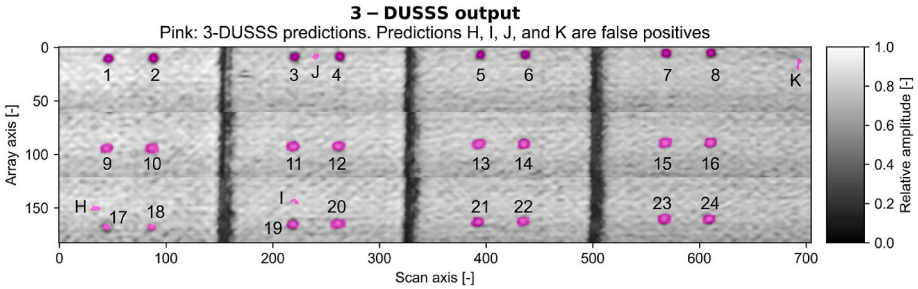

Fig. 9. 3-DUSSS segmentation output (pink) superimposed on the C-scan image of Sample A. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

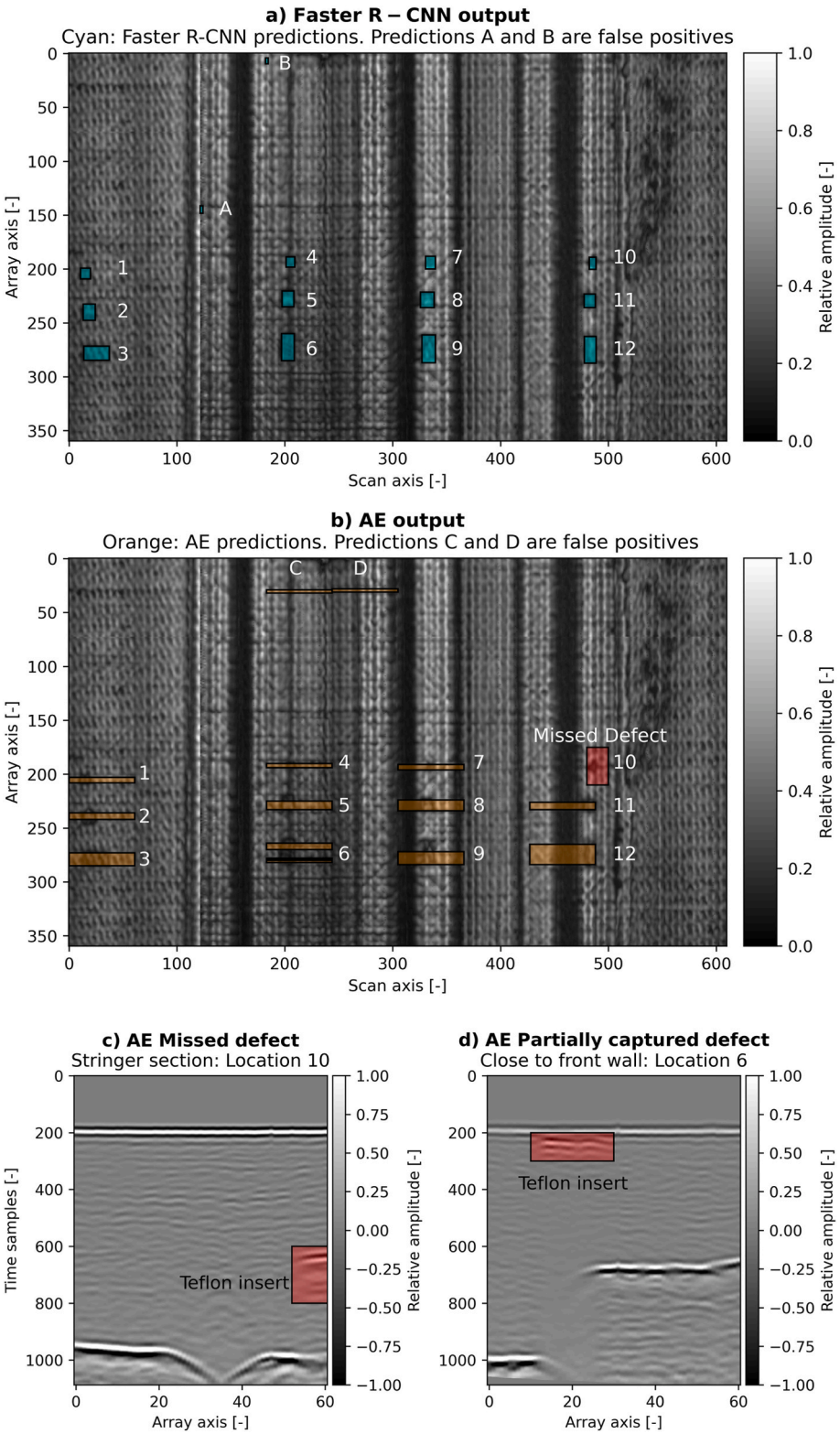

Fig. 10. A) Output of the Faster R-CNN model and b) Output of the AE model overlaid on C-scan view of the sample showing detected/missed defects (cyan and orange/red); c) Missed defect in stringer section; d) Partially captured defect close to the front wall. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

specific areas of disagreement, efficiently identifying the remaining two defects while filtering out false positives. For sample B, the models reached agreement on 11 out of 12 defects, with the human in the loop triggered to review five areas where the models disagreed, as shown in Fig. 13.

4.3. *Level 3 – high automation*

The multi model Level 3 automation produced results consistent with Level 2 in terms of agreement between the AE and Faster R-CNN models, with the key difference being that disagreements between the models

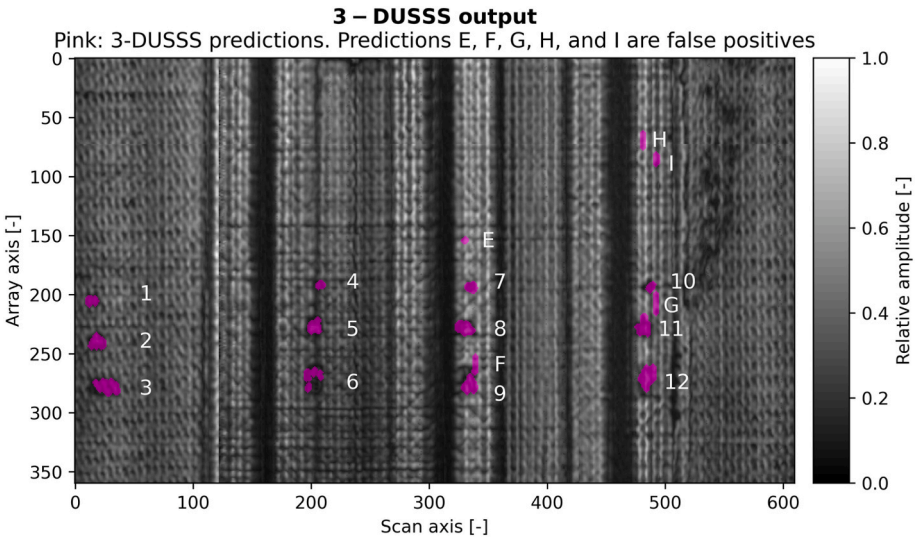

Fig. 11. 3-DUSSS segmentation output (pink) overlaid on the C-scan view of the sample B. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

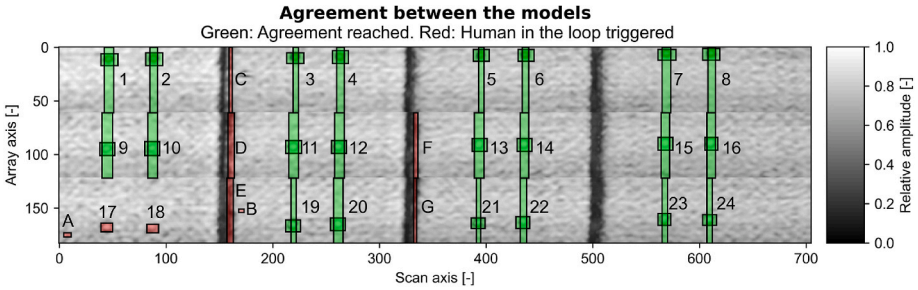

Fig. 12. Sample A) Agreement (green) and disagreement (red) between the Faster R-CNN and AE models. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

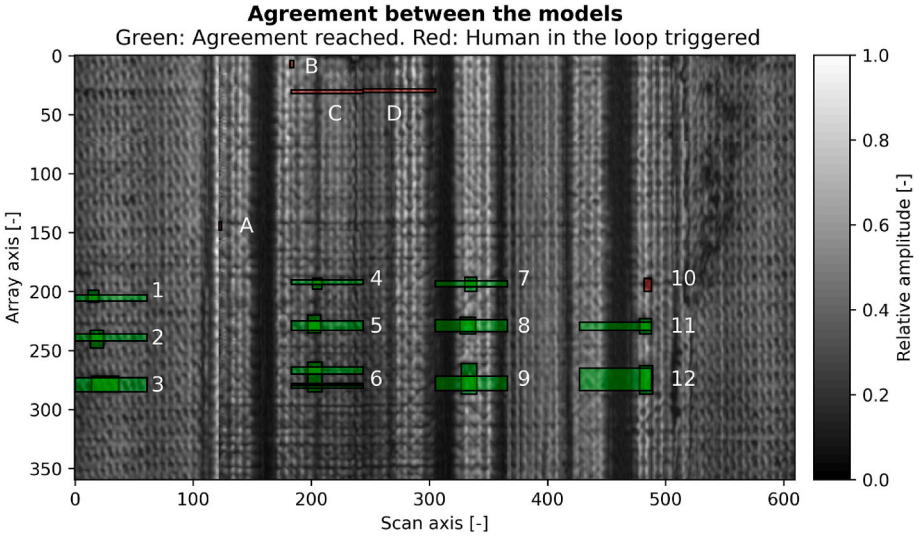

Fig. 13. Sample B: Agreement (green) and disagreement (red) between the Faster R-CNN and AE models. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

were resolved by the 3-DUSSS model rather than through a human-in-the-loop mechanism. In Sample A, the 3-DUSSS model confirmed that the two false negative calls by AE were defects, resulting in the successful identification of all 24 defects while discarding other false positive calls. An example of a disagreement in Sample A is shown in Fig. 14 a), where two defects near the back wall were detected by the Faster R-CNN but missed by the AE model (refer to Fig. 12). In this section, both the Faster R-CNN and 3-DUSSS models identified one false

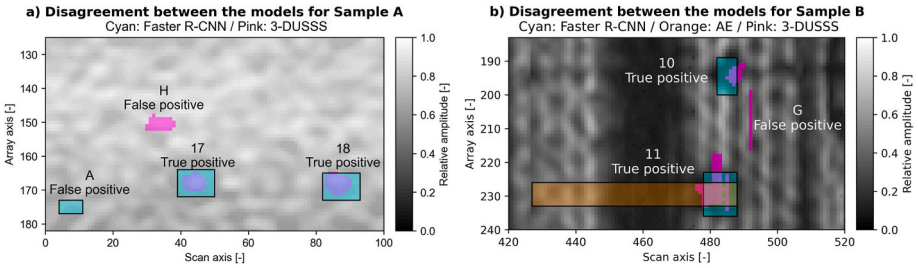

Fig. 14. Areas of disagreement between models resolved by 3-DUSSS; a) Sample A with Faster R-CNN (cyan) and 3-DUSSS (pink) predictions overlaid on the C-scan; b) Sample B with Faster R-CNN (cyan), AE (orange), and 3-DUSSS (pink) predictions overlaid on the C-scan. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

positive each. These detections were rechecked for agreement and ultimately rejected as false positives. In contrast, the two detections with coinciding results were confirmed as true positives, resolving the disagreement between the models. Inference time for the 3-DUSSS model was significantly reduced compared to processing the full volume, taking 91.59 ± 0.83 s to resolve nine areas of disagreement. While this reduction may seem modest here, it is important to highlight that these experiments were conducted on relatively small scans and reference samples. For larger datasets, typical in industrial applications, this targeted approach would likely result in more substantial time savings.

Similar results were observed in Sample B, where all defects were correctly identified. An example of model disagreement is shown in Fig. 14 b), where a 5.0 × 5.0 mm stringer defect, missed by AE, was confirmed as a true positive by the 3-DUSSS model. As in the previous example, 3-DUSSS produced one false positive, which was rejected since it did not coincide with any other model's detection. The inference time for the 3-DUSSS model to resolve five areas of disagreement was 54.12 ± 0.74 s.

Overall, Level 3 automation offers several benefits. By combining the Faster R-CNN, AE, and 3-DUSSS models, all defects in this study were successfully detected, with the 3-DUSSS model resolving areas of disagreement and filtering out false positives. This approach ensures fast, reliable results while enabling the use of less powerful hardware. Additionally, the workflow reduces both analysis time and operator workload, while still allowing operators to review intermediate results, examine areas of disagreement, and intervene if needed, thereby preventing potential performance degradation. This workflow achieves results close to the ideal fully automated process, with minimal impact on analysis time and system failure risk. An overview of the performance metrics for individual models, including the number of false positive and false negative calls, as well as inference times, is provided in Table 4. Recall is defined as the number of true positives divided by the sum of true positives and false negatives, while precision is the number of true positives divided by the sum of true and false positives. The F1 score is calculated as the harmonic mean of precision and recall.

5. Conclusion

In this paper, AI-aided data analysis strategies were explored across three proposed levels of automation, focusing on the use of multiple AI models to simultaneously process different ultrasonic views. A case study was conducted on two defective CFRP reference samples containing 36 manufactured defects. These samples were inspected using an industrial manipulator and a PAUT roller probe to simulate industrial practices for inspecting large composite components. Integrating multiple models within the NDE data analysis workflow provided flexibility

Table 4
Overview of reported performance metrics for different automation levels.

Automation Level 1	Metric	System					
		Anomaly detection AE		Faster R-CNN		3-DUSS	
		Sample A	Sample B	Sample A	Sample B	Sample A	Sample B
Single model system Operator reviews all data and all AI model outputs	Inference [s]	2.28 ± 0.12	2.34 ± 0.11	0.22 ± 0.06	0.55 ± 0.08	221.34 ± 1.41	379.98 ± 1.21
	False positives [−]	5	2	2	2	4	5
	False negatives [−]	2	1	0	0	0	0
	Precision [−]	0.815	0.846	0.923	0.857	0.857	0.706
	Recall [−]	0.917	0.917	1.000	1.000	1.000	1.000
	F1 [-]	0.863	0.880	0.960	0.923	0.923	0.828

Automation Level 2	Metric	System	
		Anomaly detection AE | Faster R-CNN	
		Sample A	Sample B
Two model system Human-in-the-loop mechanism triggered for disagreements	Inference [s]	2.44 ± 0.18 (2.28 ± 0.12 | 0.22 ± 0.06)	2.89 ± 0.19 (2.34 ± 0.11 | 0.55 ± 0.08)
	False positives [−]	7	4
	False negatives [−]	0	0
	Flagged for Human-in-the-loop mechanism [−]	9 (7 false positives and 2 true positives)	5 (4 false positives and 1 true positive)

Automation Level 3	Metric	System	
		Anomaly detection AE | Faster R-CNN | 3-DUSS	
		Sample A	Sample B
Three model system Operator moved to supervisory role	Inference [s]	94.03 ± 1.01 (2.28 ± 0.12 | 0.22 ± 0.06 | 91.59 ± 0.83)	57.01 ± 0.93 (2.34 ± 0.11 | 0.55 ± 0.08 | 54.12 ± 0.74)
	False positives [−]	0	0
	False negatives [−]	0	0

in designing workflows, managing intermediate results, and resolving model disagreements. This approach also facilitated a more robust setup leading to the successful detection of all examined defects. The study revealed that for.

- Level 1 - Operator Assistance: The conservative use of AI models prioritises safety by minimising false negatives, at the cost of increasing false positives. The suggestions provided by the models accelerate data analysis while maintaining minimal risks associated with reliance on automation. Human operators validate all AI outputs and retain full control over decision-making, resulting in only a slight increase in system complexity.
- Level 2 – Multi Model Partial Automation: Improved results were achieved by comparing outputs from two models and prompting human operators to intervene in areas of disagreement. This comparison acts as an additional validation step for reported detections, aiming to increase trust in the automated process. While this approach speeds up data analysis, it requires a higher degree of trust in the models.
- Level 3 – Multi Model Fully Automation: Incorporating the 3-DUSSS model as an arbiter enabled a simultaneous analysis workflow that processes ultrasonic B-scans, C-scans, and full volumetric data. The deployment of 3-DUSSS to only areas of disagreement greatly reduced inference times and memory requirements, making this strategy deployable on less powerful hardware. The combination of three models achieved near-ideal results while addressing model trust concerns with two layers of validation.

While this research provides an analysis of the performance of different automation levels on fabricated defects of known size and shape, there is an opportunity to explore the system's functionality when applied to naturally occurring defects, such as porosities, to assess the robustness of individual models across a wider range of defect types. Additionally, optimisation of models in terms of hyperparameter tuning, changes in architectures, or training regimes with new and varied data is deemed promising for achieving improved results.

In future work, the developed system will be integrated into a production-level industrial use case to assess its scalability, robustness, and performance in a complex real-world environment, while also addressing integration challenges with existing workflows. Additionally, the expansion of defect detection models to include a broader range of defect types, such as porosities or foreign object inclusions, will be explored.

CRediT authorship contribution statement

Vedran Tunukovic: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Shaun McKnight:** Writing – review & editing, Software, Methodology, Data curation. **Amine Hifi:** Writing – review & editing, Software. **Ehsan Mohseni:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Formal analysis. **S. Gareth Pierce:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition. **Randika K.W. Vithanage:** Writing – review & editing, Supervision, Resources. **Gordon Dobie:** Writing – review & editing, Supervision. **Charles N. MacLeod:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition. **Sandy Cochran:** Supervision, Project administration, Funding acquisition. **Tom O'Hare:** Writing – review & editing, Supervision, Investigation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gareth Pierce reports financial support was provided by Spirit Aero Systems Inc. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported through EPSRC Centre for Doctoral Training in Future Ultrasonic Engineering (FUSE CDT) EP/S023879/1, and Spirit AeroSystems/Royal Academy of Engineering Research Chair for In-Process Non-Destructive Testing of Composites, RCSRF 1920/10/32.

Data availability

The authors do not have permission to share data.

References

[1] Mineo C, Pierce SG, Nicholson PI, Cooper I. Robotic path planning for non-destructive testing – a custom MATLAB toolbox approach. Robot. Comput.-Integr. Manuf. Feb. 2016;37:1–12. https://doi.org/10.1016/j.rcim.2015.05.003.
[2] Mineo C, Pierce SG, Wright B, Nicholson PI, Cooper I. Robotic path planning for non-destructive testing of complex shaped surfaces. AIP Conf Proc Mar. 2015;1650(1):1977–87. https://doi.org/10.1063/1.4914825.
[3] Mineo C, et al. Flexible integration of robotics, ultrasonics and metrology for the inspection of aerospace components. AIP Conf Proc Feb. 2017;1806(1). https://doi.org/10.1063/1.4974567. 020026–020026.
[4] Mangalgiri PD. Composite materials for aerospace applications. Bull Mater Sci 1999;22(3):657–64.
[5] Bachmann J, Hidalgo C, Bricout S. Environmental analysis of innovative sustainable composites with potential use in aviation sector—a life cycle assessment review. Sci China Technol Sci Sep. 2017;60(9):1301–17. https://doi.org/10.1007/S11431-016-9094-Y.
[6] Schnars U, Henrich R. 'Applications of NDT methods on composite structures in aerospace industry', presented at the conference on damage in composite materials. 2006.
[7] Kapadia A. National composites network best practice guide non destructive testing of composite materials [Online]. Available: http://www.twi.co.uk/j32k/index.xtp; 2007.
[8] Wooh SC, Shi Y. Optimum beam steering of linear phased arrays. Wave Motion 1999;29(3):245–65. https://doi.org/10.1016/S0165-2125(98)00039-0.
[9] Holmes C, Drinkwater BW, Wilcox PD. Post-processing of the full matrix of ultrasonic transmit–receive array data for non-destructive evaluation. NDT E Int Dec. 2005;38(8):701–11. https://doi.org/10.1016/J.NDTEINT.2005.04.002.
[10] Wilcox PD. Ultrasonic arrays in NDE: beyond the B-scan. AIP Conf Proc Jan. 2013;1511(1). https://doi.org/10.1063/1.4789029. 33–33.
[11] Duernberger E, MacLeod C, Lines D, Loukas C, Vasilev M. Adaptive optimisation of multi-aperture ultrasonic phased array imaging for increased inspection speeds of wind turbine blade composite panels. NDT E Int Dec. 2022;132:102725. https://doi.org/10.1016/j.ndteint.2022.102725.
[12] Mineo C. Automated NDT inspection for large and complex geometries of composite materials. 2015. https://doi.org/10.48730/GXQ8-WA04.
[13] Drai R, Sellidj F, Khelil M, Benchaala A. Elaboration of some signal processing algorithms in ultrasonic techniques: application to materials NDT. Ultrasonics 2000;38:503–7.
[14] Barut S, Bissauge V, Ithurralde G, Claassens W. Computer-aided analysis of ultrasound data to speed-up the release of aerospace CFRP components. presented at the 18th World Conference on Nondestructive Testing, Durban, South Africa: e-Journal of Nondestructive Testing Apr. 2012;17(7).
[15] Aldrin JC, Coughlin C, Forsyth DS, Welter JT. Progress on the development of automated data analysis algorithms and software for ultrasonic inspection of composites. AIP Conf Proc Feb. 2014;1581(1):1920–7. https://doi.org/10.1063/1.4865058.
[16] Cantero-Chinchilla S, Wilcox PD, Croxford AJ. Deep learning in automated ultrasonic NDE – developments, axioms and opportunities. Dec. 2021. https://doi.org/10.48550/arxiv.2112.06650. ArXiv211206650 Eess.
[17] McKnight S, et al. A comparison of methods for generating synthetic training data for domain adaption of deep learning models in ultrasonic non-destructive evaluation. NDT E Int Jan. 2024;141:102978. https://doi.org/10.1016/j.ndteint.2023.102978.
[18] Mery D. Aluminum casting inspection using deep object detection methods and simulated ellipsoidal defects. Mach Vis Appl Apr. 2021;32(3):72. https://doi.org/10.1007/s00138-021-01195-5.
[19] A Computational Framework for Automatic Online Path Generation of Robotic Inspection Tasks via Coverage Planning and Reinforcement Learning | IEEE Journals & Magazine | IEEE Xplore'. Accessed: October. 16, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8476296.
[20] Pauli V, Leinonen J. Technology survey on NDT of carbon-fiber composites [Online]. Available: https://www.theseus.fi/bitstream/handle/10024/54515/vaa

ra%20leinonen%20B%208%202012.pdf?sequence=1. [Accessed 31 October 2023].
[21] Munir N, Park J, Kim HJ, Song SJ, Kang SS. Performance enhancement of convolutional neural network for ultrasonic flaw classification by adopting autoencoder. NDT E Int Apr. 2020;111. https://doi.org/10.1016/J.NDTEINT.2020.102218. 102218–102218.
[22] Amiri N, Farrahi GH, Kashyzadeh KR, Chizari M. Applications of ultrasonic testing and machine learning methods to predict the static & fatigue behavior of spot-welded joints. J Manuf Process Apr. 2020;52:26–34. https://doi.org/10.1016/J.JMAPRO.2020.01.047.
[23] Meng M, Chua YJ, Wouterson E, Ong CPK. Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks. Neurocomputing Sep. 2017;257:128–35. https://doi.org/10.1016/J.NEUCOM.2016.11.066.
[24] Guo Y, et al. Fully convolutional neural network with GRU for 3D braided composite material flaw detection. IEEE Access 2019;7:151180–8. https://doi.org/10.1109/ACCESS.2019.2946447.
[25] Cheng X, Ma G, Wu Z, Zu H, Hu X. Automatic defect depth estimation for ultrasonic testing in carbon fiber reinforced composites using deep learning. NDT E Int Apr. 2023;135:102804. https://doi.org/10.1016/j.ndteint.2023.102804.
[26] Medak D, Posilovic L, Subasic M, Budimir M, Loncaric S. Automated defect detection from ultrasonic images using deep learning. IEEE Trans Ultrason Ferroelectrics Freq Control Oct. 2021;68(10):3126–34. https://doi.org/10.1109/TUFFC.2021.3081750.
[27] Medak D, Posilovic L, Subasic M, Budimir M, Loncaric S. Deep learning-based defect detection from sequences of ultrasonic B-scans. IEEE Sens J Feb. 2022;22(3):2456–63. https://doi.org/10.1109/JSEN.2021.3134452.
[28] Medak D, Posilović L, Subašić M, Budimir M, Lončarić S. DefectDet: a deep learning architecture for detection of defects with extreme aspect ratios in ultrasonic images. Neurocomputing Feb. 2022;473:107–15. https://doi.org/10.1016/J.NEUCOM.2021.12.008.
[29] Milković F, Filipović B, Subašić M, Petković T, Lončarić S, Budimir M. Ultrasound anomaly detection based on variational autoencoders. In: 2021 12th international symposium on image and signal processing and analysis (ISPA); Sep. 2021. p. 225–9. https://doi.org/10.1109/ISPA52656.2021.9552041.
[30] Posilović L, Medak D, Milković F, Subašić M, Budimir M, Lončarić S. Deep learning-based anomaly detection from ultrasonic images. Ultrasonics Aug. 2022;124. https://doi.org/10.1016/J.ULTRAS.2022.106737. 106737–106737.
[31] Li C, et al. Intelligent damage recognition of composite materials based on deep learning and ultrasonic testing. AIP Adv Dec. 2021;11(12):125227. https://doi.org/10.1063/5.0063615.
[32] Yunker A, Lake R, Kettimuthu R, Kral Z. Comparative study on deep learning methods for defect identification and classification in composite aerostructure material. Presented at the 2023 50th annual review of progress in quantitative nondestructive evaluation. American Society of Mechanical Engineers Digital Collection; Jul. 2023. https://doi.org/10.1115/QNDE2023-108602.
[33] Trouvé-Peloux P, Abeloos B, Ben Fekih A, Trottier C, Roche J-M. Benefit of neural network for the optimization of defect detection on composite material using ultrasonic non destructive testing. Presented at the 2021 48th annual review of progress in quantitative nondestructive evaluation. American Society of Mechanical Engineers Digital Collection; Jan. 2022. https://doi.org/10.1115/QNDE2021-75925.
[34] Posilović L, Medak D, Subašić M, Budimir M, Lončarić S. Generating ultrasonic images indistinguishable from real images using Generative Adversarial Networks. Ultrasonics Feb. 2022;119. https://doi.org/10.1016/J.ULTRAS.2021.106610. 106610–106610.
[35] Virkkunen I, Koskinen T, Jessen-Juhler O, Rinta-aho J. Augmented ultrasonic data for machine learning. J Nondestruct Eval Mar. 2021;40(1):1–11. https://doi.org/10.1007/S10921-020-00739-5/TABLES/1.
[36] De Fauw J, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. Sep. 2018;24(9):1342–50. https://doi.org/10.1038/s41591-018-0107-6.
[37] Zhou W, et al. Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images. Nat Commun Feb. 2021;12(1):1259. https://doi.org/10.1038/s41467-021-21466-z.
[38] Bertovic M. A human factors perspective on the use of automated aids in the evaluation of NDT data. AIP Conf Proc Feb. 2016;1706(1). https://doi.org/10.1063/1.4940449. 020003–020003.
[39] Vagia M, Transeth AA, Fjerdingen SA. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? Appl Ergon Mar. 2016;53:190–202. https://doi.org/10.1016/j.apergo.2015.09.013.
[40] Bertovic M, Virkkunen I. Nde 4.0: new paradigm for the NDE inspection personnel. Handb. Nondestruct. Eval 2021;40:1–31. https://doi.org/10.1007/978-3-030-48200-8_9-1.
[41] Endsley MR, Kiris EO. The out-of-the-loop performance problem and level of control in automation. Hum Factors Jun. 1995;37(2):381–94. https://doi.org/10.1518/001872095779064555.
[42] Parasuraman R, Mouloua M, Molloy R. Effects of adaptive task allocation on monitoring of automated systems. Hum Factors Dec. 1996;38(4):665–79. https://doi.org/10.1518/001872096778827279.
[43] Mui L, Mohtashemi M, Halberstadt A. Notions of reputation in multi-agents systems: a review. In: *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, in AAMAS '02. New York, NY, USA: Association for Computing Machinery; Jul. 2002. p. 280–7. https://doi.org/10.1145/544741.544807.
[44] Yin M, Wortman Vaughan J, Wallach H. Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 CHI conference on human factors in computing systems, in CHI '19. New York, NY, USA: Association for Computing Machinery; May 2019. p. 1–12. https://doi.org/10.1145/3290605.3300509.
[45] Bravo-Rocca G, Liu P, Guitart J, Dholakia A, Ellison D, Hodak M. Human-in-the-loop online multi-agent approach to increase trustworthiness in ML models through trust scores and data augmentation. May 02, 2022. https://doi.org/10.48550/arXiv.2204.14255. *arXiv*: arXiv:2204.14255.
[46] Kore A. Designing human-centric AI experiences: applied UX design for artificial intelligence. In: Design thinking. Berkeley, CA: Apress; 2022. https://doi.org/10.1007/978-1-4842-8088-1.
[47] Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the Predictions of Any Classifier' Aug. 09, 2016. https://doi.org/10.48550/arXiv.1602.04938. *arXiv*: arXiv:1602.04938.
[48] Pyle RJ, Hughes RR, Wilcox PD. Interpretable and explainable machine learning for ultrasonic defect sizing. IEEE Trans Ultrason Ferroelectrics Freq Control Apr. 2023;70(4):277–90. https://doi.org/10.1109/TUFFC.2023.3248968.
[49] Olympus-ims. RollerFORM: phased array wheel probe manual [Online]. Available: https://www.olympus-ims.com/en/rollerform/; 2023.
[50] MicoPulse 6PAPhased Array Ultrasonic TechnologyPeak NDT, [Online]. Available: https://www.peakndt.com/products/micropulse-6pa/.
[51] Schunk. SCHUNK Force Torque sensors manual [Online]. Available: https://schunk.com/us/en/automation-technology/force/torque-sensors/ft/ftn-gamma-si-130-10/p/EPIM_ID-30865; 2023.
[52] KUKA Robotics. KUKA KR90 R3100 extra HA specification manual [Online]. Available: https://www.kuka.com/-/media/kuka-downloads/imported/8350ff3ca11642998dbdc81dcc2ed44c/0000208694_en.pdf; 2023.
[53] Blain P, et al. In: Artificial defects in CFRP composite structure for thermography and shearography nondestructive inspection, vol. 10449; Jun. 2017. p. 562–71. https://doi.org/10.1117/12.2271701. 13.
[54] Paszke A, et al. Automatic differentiation in PyTorch [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ. [Accessed 8 July 2024].
[55] Tunukovic V, et al. Unsupervised machine learning for flaw detection in automated ultrasonic testing of carbon fibre reinforced plastic composites. Ultrasonics May 2024;140:107313. https://doi.org/10.1016/j.ultras.2024.107313.
[56] Virtanen P, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods Mar. 2020;17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2.
[57] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, 'A density-based algorithm for discovering clusters in large spatial databases with noise'.
[58] Tunukovic V, et al. A study of machine learning object detection performance for phased array ultrasonic testing of carbon fibre reinforced plastics. NDT E Int Jun. 2024;144:103094. https://doi.org/10.1016/j.ndteint.2024.103094.
[59] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell Jun. 2015;39(6):1137–49. https://doi.org/10.48550/arxiv.1506.01497.
[60] McKnight S, et al. 3-DUSSS: 3-dimensional ultrasonic self supervised segmentation. arXiv: arXiv:2411.07835 2024;12. https://doi.org/10.48550/arXiv.2411.07835. Nov.